

Complete sequence of the human RNA polymerase II largest subunit

Marguerite Wintzerith, Joel Acker, Serge Vicaire, Marc Vigneron and Claude Kedinger
Laboratoire de Génétique Moléculaire des Eucaryotes (CNRS), Unité 184 de Biologie Moléculaire et de Génétique (INSERM), 11 rue Humann, 67085 Strasbourg Cedex, France

Submitted January 21, 1992 EMBL accession no. X63564

Eukaryotic RNA polymèrase II (or B) consists of 10 to 14 polypeptides ranging from 220 to 10 kD (1). The understanding in molecular terms of the initiation and elongation processes and their control by specific trans-acting regulatory proteins will ultimately require the cloning of all components of the transcriptional apparatus. We describe the molecular cloning and characterization of the cDNA encoding the human polymerase II largest subunit, RPBh1. An earlier attempt had led to the isolation of a partial human genomic clone (2).

In the present study, lambda-Zap HeLa cell cDNA libraries were screened with probes derived by specific PCR-mediated amplification of mouse cDNA sequences (3). Several overlapping cDNA clones were isolated and sequenced on each DNA strand by automated or manual sequencing. One of them comprised the entire RPBh1 coding region. The predicted 1970 amino acid sequence (calculated MW of 217205), is presented in Fig. 1. The comparison of the human sequence with the corresponding published mouse sequence (3) revealed two differences. The first alteration is a threonine residue at position 1856 in man, instead of an alanine in mouse. This change occurs within the conserved carboxy-terminal domain (CTD) of the protein, on the 38th heptapeptide repeat YTPTSPK, as originally defined (3). Note that the sequence of this element in hamster (4) is identical to that in man. The second difference is the sequence between residues 1498 and 1537, which replaces a single arginine residue in the mouse homologue. This element most likely corresponds to an additional exon which must have been overlooked upon reconstruction of the predicted coding frame from the mouse genomic sequence. Reexamination of the sequence of intron 26 (see 3 and EMBL M12130) indeed revealed motifs compatible with genuine splice acceptor (5′-CTCTTTGCAG-3′) and donor (5′-GTGAGT-3′) sites, on the 5′ and 3′ sides of this exon. If one corrects the mouse sequence for this exon, the alanine to threonine alteration in the CTD remains the only difference between the mouse and human proteins, with a corresponding nucleotide conservation of 89.4%. Strikingly, the 5′ untranslated region is significantly more conserved (about 85%) than the 3′ untranslated region (about 50%) in these two organisms.

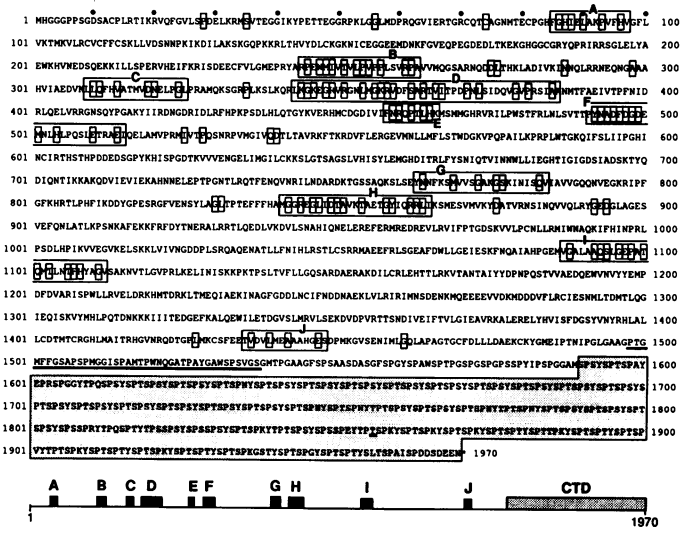
The results of a computer analysis of homologous peptide sequences from prokaryotes and eukaryotes, are summarized in Fig. 1. A number of invariant residues were identified, most of which delineate conserved domains (A–J). Although the high degree of conservation of these domains may reflect functional significances, their role remains unknown. The particularly elevated number of basic residues in domain D suggests its involvement in a possible DNA-binding region. A putative C_2H_2-type zinc-binding element (residues 71 to 87), conserved among eukaryotes, may also contribute to this function (3) or be implicated in important protein-protein interactions. Clearly, these conserved domains constitute privileged targets for future functional analyses.

ACKNOWLEDGEMENTS

We thank C.Hauss for technical assistance, J.M.Garnier for providing cDNA libraries, the chemistry staff for oligonucleotides and C. Werlé for artwork. This work was supported by the Association pour la Recherche sur le Cancer. J.A. was a fellow of the Ligue Nationale contre le Cancer.

REFERENCES

1. Sawadogo,M. and Sentenac,A. (1990) *Annu. Rev. Biochem.* **59**, 711–754.
2. Cho,K.W.Y. *et al.* (1985) *J. Biol. Chem.* **260**, 15204–15210.
3. Ahearn,J.M. *et al.* (1987) *J. Biol. Chem.* **262**, 10695–10705.
4. Allison,L.A. *et al.* (1988) *Mol. Cell. Biol.* **8**, 321–329.
5. Jockerst,R.S. *et al.* (1989) *Mol. Gen. Genet.* **215**, 266–275.
6. Allison,L.A. *et al.* (1985) *Cell* **42**, 599–610.
7. Pühler,G. *et al.* (1989) *Nucleic Acids Res.* **17**, 4517–4534.
8. Leffers,H. *et al.* (1989) *J. Mol. Biol.* **206**, 1–17.
9. Ovchinnikov,Y.A. *et al.* (1982) *Nucleic Acids Res.* **10**, 4035–4044.



Predicted peptide sequence of the RPBh1 cDNA. Amino acid residues underlined are discussed in the text. Those which are unchanged in *H. sapiens*, *M. musculus* (3), *D. melanogaster* (5), *S. cerevisiae* (6), *S. acidocaldarius* (7), *H. halobium* (8) and *E. coli* (9) are boxed. Unvariant residues separated by 5 or less variable residues have been arbitrarily grouped as conserved domains (A–J). The carboxy-terminal domain (CTD) spanning the 52 heptapeptide repeats is shadowed. A schematic representation of RPBh1 is shown at the bottom, with the relative positions of the A–J and CTD domains drawn at scale.